The Analysis of Data from Continuous Probability Distributions

Timothy E. Holy
Department of Physics, Princeton University, Princeton, New Jersey, 08544
(June 10, 1997)

Abstract

Conventional statistics begins with a model, and assigns a likelihood of obtaining any particular set of data. The opposite approach, beginning with the data and assigning a likelihood to any particular model, is explored here for the case of points drawn randomly from a continuous probability distribution. A scalar field theory is used to assign a likelihood over the space of probability distributions. The most likely distribution may be calculated, providing an estimate of the underlying distribution and a convenient graphical representation of the raw data. Fluctuations around this maximum likelihood estimate are characterized by a robust measure of goodness-of-fit. Its distribution may be calculated by integrating over fluctuations. The resulting method of data analysis has some advantages over conventional approaches.

When the outcome of an experiment falls into one of a few categories, the frequency of a particular outcome is an estimate of its probability. For example, by repeatedly flipping a coin we learn about the probability of obtaining heads. But when the outcome of an experiment is one of a continuum, no finite set of data can determine the frequency of each outcome. One common method of estimating the underlying probability distribution is to group observations into categories, a procedure known as "binning." The histogram (the frequency of observations in each bin) is then used as an estimate of the underlying probability distribution. While binning is widely used, it has a number of undesirable consequences. It requires a choice of bins (both their number and sizes), and different choices lead to different histograms. Thus even the appearance of raw data, when presented in graphical format, depends on arbitrary choices. Binning also throws information away, since different outcomes are grouped together.

An alternative approach has been presented [1,2] to estimate the probability distribution. These authors assign a likelihood $P[Q|x_1,\ldots,x_N]$ that the distribution $Q(x)$ describes the data $x_1,\ldots,x_N$. The underlying distribution might then be estimated as the one which maximizes $P[Q|x_1,\ldots,x_N]$. By Bayes' rule,

$$P[Q|x_1,\ldots,x_N] = \frac{P[x_1,\ldots,x_N|Q]P[Q]}{P[x_1,\ldots,x_N]} \tag{1}$$

$$= \frac{Q(x_1)\cdots Q(x_N)P[Q]}{\int \mathcal{D}Q\, Q(x_1)\cdots Q(x_N)P[Q]}, \tag{2}$$

where $P[Q]$ is some *a priori* likelihood of the distribution Q. As no finite set of data can specify an arbitrary function of a continuous variable, a choice for $P[Q]$ is necessary to regularize the inverse problem. This choice encapsulates our baises in an explicit fashion. (These biases are implicit in other approaches, e.g., in our interpretation of a histogram.)

What form should $P[Q]$ have? By setting $Q(x) = \psi^2(x)$ [1], where ψ may take any value in $(-\infty, \infty)$, we may insure that Q is non-negative. ψ will be referred to as the *amplitude* by analogy with quantum mechanics. $P[Q]$ should incorporate our bias that Q be "smooth" [3]. "Smoothness" is enforced by penalizing large gradients in Q—or rather, in ψ. Finally, Q should be normalized. In one dimension, the *a priori* distribution is

$$P[\psi] = \frac{1}{Z} \exp\left[-\int dx\, \frac{\ell^2}{2}(\partial_x \psi)^2\right] \delta\left(1 - \int dx\, \psi^2\right), \tag{3}$$

where Z is the normalization factor and ℓ is a constant which controls the penalty applied to gradients. The delta function enforces normalization of the distribution Q.

The probability $P[Q|x_1, \ldots, x_N]$ of a distribution Q, given the data, is therefore

$$\begin{aligned} P[\psi|x_1, \ldots, x_N] &\propto \psi^2(x_1) \cdots \psi^2(x_N) \\ &\quad \times \exp\left[-\int dx\, \frac{\ell^2}{2}(\partial_x \psi)^2\right] \delta\left(1 - \int dx\, \psi^2\right) \qquad (4) \\ &= e^{-S[\psi]} \delta\left(1 - \int dx\, \psi^2\right), \qquad (5) \end{aligned}$$

where the effective action S is

$$S[\psi] = \int dx\, \left(\frac{\ell^2}{2}(\partial_x \psi)^2 - 2 \ln \psi \sum_i \delta(x - x_i)\right). \tag{6}$$

What is the most likely distribution (amplitude), given the data? From Eq. (5), this is the ψ which minimizes the action, subject to the normalization constraint. This ψ will be called the classical amplitude, $\psi_{\rm cl}$. To handle the normalization constraint, we subtract a Lagrange multiplier term $\lambda(1 - \int dx\, \psi^2)$ from the action; $\psi_{\rm cl}$ satisfies the equations

$$-\ell^2 \partial_x^2 \psi_{\rm cl} + 2\lambda \psi_{\rm cl} - \frac{2}{\psi_{\rm cl}} \sum_i \delta(x - x_i) = 0, \tag{7a}$$

$$\int dx\, \psi_{\rm cl}^2 = 1. \tag{7b}$$

The solution to these equations may be written

$$\psi_{\rm cl}(x) = \sqrt{\kappa} \sum_i a_i e^{-\kappa |x - x_i|}, \tag{8}$$

where $\kappa^2 = 2\lambda/\ell^2$. Each data point therefore contributes one peak of width $1/\kappa$ to the amplitude $\psi_{\rm cl}$. This is reminiscent of kernel estimation [4], using the amplitude rather than the probability distribution. Eqs. (7) imply

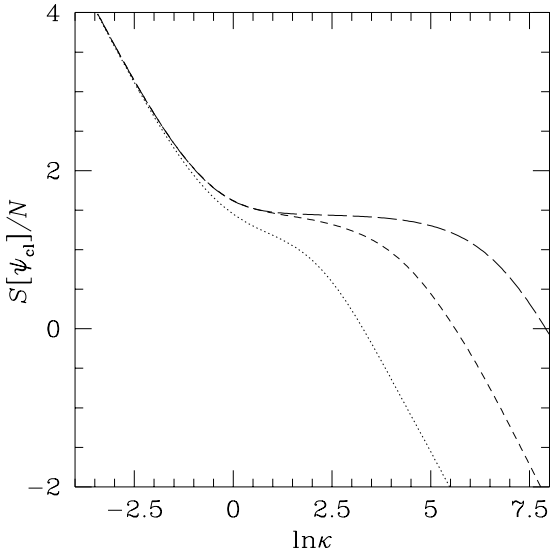

FIG. 1. The classical action, Eq. (10), as a function of $\ln\kappa$ for data drawn randomly from a gaussian distribution with zero mean and unit variance. Long dash, $N = 2000$; short dash, $N = 200$; dots, $N = 20$.

$$2\lambda a_i \sum_j a_j e^{-\kappa|x_i - x_j|} = 1, \qquad i = 1, \ldots, N \tag{9a}$$

$$\frac{N}{2\lambda} + \sum_{i,j} a_i a_j \, \kappa|x_i - x_j| \, e^{-\kappa|x_i - x_j|} = 1. \tag{9b}$$

These $N + 1$ equations determine λ and the a_i as a function of κ [5].

Using the equation of motion, Eqs. (7), the classical action $S[\psi_{\rm cl}]$ may be written

$$S[\psi_{\rm cl}] = N - \lambda(\kappa) - \sum_i \ln Q_{\rm cl}(x_i). \tag{10}$$

For the proper choice of κ one might hope that $Q_{\rm cl} \approx \bar{Q}$, the true distribution. Since the data points x_i arise from the true distribution $\bar{Q}(x)$, we expect

$$\sum_i \delta(x - x_i) \approx N\bar{Q}(x). \tag{11}$$

Therefore, the last term of Eq. (10) is approximately $N \int dx \, \bar{Q}(x) \ln \bar{Q}(x)$, which can be interpreted as the entropy (or the information [6]). Using perturbation theory one may show that when $Q_{\rm cl} \approx \bar{Q}$, then $\lambda \approx N$, so the first two terms of Eq. (10) (the penalty for gradients) approximately cancel (more precisely, increase much less rapidly than N).

How does one choose κ? In Figure 1, the classical action is plotted against $\ln\kappa$ for data sets generated from a gaussian distribution. One sees that, over a region of width $\ln N$, $S[\psi_{\rm cl}]$ is insensitive to the precise choice of κ. Therefore, κ may be chosen by finding the point of minimum sensitivity $|dS[\psi_{\rm cl}]/d\ln\kappa|$ [7,8].

Once κ has been chosen, the maximum likelihood distribution $Q_{\rm cl}(x) = \psi^2_{\rm cl}(x)$ is uniquely determined. An example of results from this procedure are shown in Figure 2. One sees

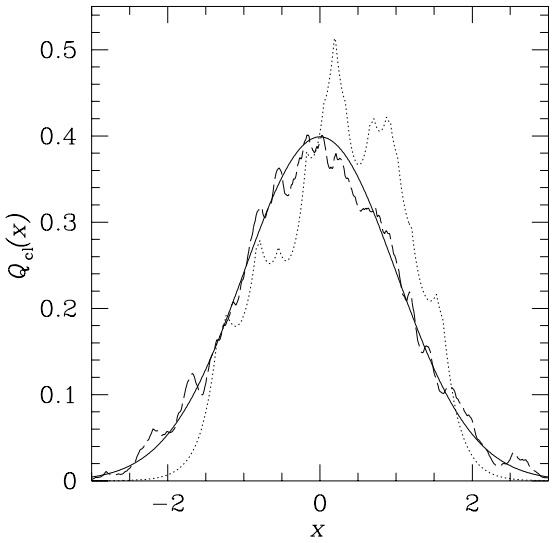

FIG. 2. The classical distribution $Q_{\rm cl}$, for data drawn randomly from a gaussian distribution (solid line). Dashed curve, $N = 2000$; dotted curve, $N = 20$.

convergence towards the underlying distribution as N increases. Note that even for $N = 20$ the estimate $Q_{\rm cl}$ is illuminating; the advantages of this method over binning are especially great for small data sets.

While $Q_{\rm cl}$ represents the most likely distribution, other "nearby" distributions should also be considered. The action may be expanded around the classical amplitude, which to second order in the fluctuations $\delta\psi$ yields [9]

$$S[\psi_{\rm cl} + \delta\psi] \approx S[\psi_{\rm cl}] + \frac{1}{4}\chi^2[\delta\psi] + \int dx \left(\frac{\ell^2}{2}(\partial_x \delta\psi)^2 + \lambda\delta\psi^2\right), \tag{12}$$

where

$$\chi^2[\delta\psi] = 4\sum_i \frac{\delta\psi^2(x_i)}{\psi^2_{\rm cl}(x_i)}. \tag{13}$$

χ^2 is a measure of the goodness of fit between a trial distribution $Q = \psi^2$ and the data. It is the direct analogue of the conventional χ^2 (which here will be called χ^2_1); to see this, re-write χ^2 as

$$\chi^2 = 4\int dx\, \frac{(\psi(x) - \psi_{\rm cl}(x))^2}{\psi^2_{\rm cl}(x)} \sum_i \delta(x - x_i) \approx 4N \int dx \left(\sqrt{Q} - \sqrt{\bar{Q}}\right)^2 \tag{14}$$

using Eq. (11). Now suppose that Q and $\bar{Q}$ are close, $Q(x) = \bar{Q}(x) + \epsilon(x)$. Then we may expand the difference of square roots as

$$\left(\sqrt{Q}-\sqrt{\bar{Q}}\right)^2 \approx \frac{1}{4}\frac{\epsilon^2}{\bar{Q}}, \tag{15}$$

which establishes the connection to the traditional definition χ_1^2.

This definition of χ^2 has a number of advantages over χ_1^2. Because of the quadratic dependence on ϵ and the $\bar{Q}$ term in the denominator, χ_1^2 is quite sensitive to the tails of distributions. In contrast, χ^2 as defined in Eq. (13) is robust. It is linear in $|\epsilon|$ when $|\epsilon|$ is large, and has no potentially small term in the denominator. Therefore, this definition χ^2 is more robust than χ_1^2. Another advantage is that binning is unnecessary. This eliminates the problems of lost information and arbitrary bin-sizes and -boundaries (and simplifies the process of fitting, as one need not worry about shifting bin-boundaries). Finally, this definition of χ^2 is essentially symmetric (exactly so in Eq. (14)), and consequently is a true metric on the space of probability distributions. (The form in Eq. (14) is known as the squared Hellinger distance [4].)

How is χ^2 distributed? To lowest order, the likelihood of any particular fluctuation η is

$$\begin{aligned} P[\eta|x_1,\ldots,x_N] &\propto \delta\left(\int dx\,\psi_{\rm cl}\eta\right) \\ &\times \exp\left(-\frac{1}{4}\chi^2[\eta]-\int dx\,\left(\frac{\ell^2}{2}(\partial_x\eta)^2+\lambda\eta^2\right)\right). \end{aligned} \tag{16}$$

The distribution $P(\chi^2)$ may in principle be calculated by integrating Eq. (16) over all η with fixed χ^2; a realizable alternative is to calculate its Laplace transform, $\tilde{P}(\alpha)=\langle e^{-\alpha\chi^2[\eta]}\rangle$, where the expectation is relative to the distribution of η in Eq. (16).

One challenge in evaluating any integral over η is the "orthogonality condition" $\delta\left(\int dx\,\psi_{\rm cl}\eta\right)$ in Eq. (16). One way to handle this condition is to use the delta-function representation $\delta(y)=\lim_{\epsilon\to 0^+}\frac{1}{\sqrt{\pi\epsilon}}e^{-y^2/\epsilon}$. This adds a term $(\int dx\,\psi_{\rm cl}\eta)^2/\epsilon$ to the argument of the exponential; the path integral may then be expressed formally in terms of $\det(\mathbf{L}+\psi_{\rm cl}\otimes\psi_{\rm cl}/\epsilon)^{-1/2}$, where $\mathbf{L}$ is the appropriate operator (arising from the action, Eq. (12)) and $\psi_{\rm cl}\otimes\psi_{\rm cl}$ is the matrix with the (x,x') element equal to $\psi_{\rm cl}(x)\psi_{\rm cl}(x')$. The non-local terms proportional to $\frac{1}{\epsilon}$ are large and must be handled first. We know that $\lim_{\epsilon\to 0^+}\epsilon\det(\mathbf{L}+\psi_{\rm cl}\otimes\psi_{\rm cl}/\epsilon)$ must be finite, so all the terms diverging worse than $\frac{1}{\epsilon}$ in the determinant must vanish. (This happens because of the all-order singularity of the matrix $\psi_{\rm cl}\otimes\psi_{\rm cl}$.) So even though $\frac{1}{\epsilon}$ is large, we may evaluate this determinant exactly by working to first order in $\frac{1}{\epsilon}$. Therefore

$$\begin{aligned} \det\left(\mathbf{L}+\frac{\psi_{\rm cl}\otimes\psi_{\rm cl}}{\epsilon}\right) &= \det\mathbf{L}\det\left(1+\frac{\mathbf{L}^{-1}\psi_{\rm cl}\otimes\psi_{\rm cl}}{\epsilon}\right) \\ &= \det\mathbf{L}\left(1+\frac{\mathrm{Tr}(\mathbf{L}^{-1}\psi_{\rm cl}\otimes\psi_{\rm cl})}{\epsilon}\right). \end{aligned} \tag{17}$$

Now we can take the limit $\epsilon\to 0^+$; the integral over all η is now complete. The distribution of χ^2 (properly normalized) is therefore

$$\tilde{P}(\alpha)=\left[\frac{D(\gamma)T(\gamma)}{D(1)T(1)}\right]^{-1/2}, \tag{18}$$

where $\gamma = 4\alpha + 1$,

$$D(\gamma) = \frac{\det(-\ell^2\partial_x^2 + 2\lambda + 2\gamma\sum_i \delta(x - x_i)/Q_{\rm cl})}{\det(-\ell^2\partial_x^2 + 2\lambda)}, \tag{19}$$

$$T(\gamma) = \int dx\, dx'\, K_\gamma(x, x')\psi_{\rm cl}(x)\psi_{\rm cl}(x'), \tag{20}$$

and the propagator $K_\gamma = \mathbf{L}^{-1}$ satisfies

$$-\ell^2\partial_x^2 K_\gamma + 2\lambda K_\gamma + \frac{2\gamma}{Q_{\rm cl}}\sum_i \delta(x - x_i)K_\gamma = \delta(x - x'). \tag{21}$$

The terms of Eq. (18) can be evaluated exactly. First, consider the ratio of the determinants, Eq. (19). Standard techniques [10] allow one to express $D(\gamma)$ as the limit as $x \to \infty$ of the function $E(x;\gamma)$, where E satisfies

$$-\partial_x^2 E - 2\kappa\partial_x E + \frac{\gamma\kappa^2}{\lambda Q_{\rm cl}}\sum_i \delta(x - x_i)E = 0 \tag{22}$$

and $E(x) = 1$ for x smaller than the smallest data point. Between data points, $E(x) = E_i + F_i e^{-2\kappa(x-x_i)}$, and a short calculation shows that E_i and F_i satisfy a simple recursion relation.

The traces $T(\gamma)$ are computed as follows: let $g_\gamma(x) = \int dx'\, K_\gamma(x, x')\psi_{\rm cl}(x')$ and $g_0 = \int dx'\, K_0(x, x')\psi_{\rm cl}(x')$. g_γ may be parametrized as

$$g_\gamma(x) = g_0(x) + \frac{\sqrt{\kappa}}{4\lambda}\sum_i c_i e^{-\kappa|x-x_i|}, \tag{23}$$

and from Eq. (21) the c_i satisfy the linear equations

$$c_i + \gamma\mu_i \sum_j [c_j + (1 + \kappa|x_i - x_j|)a_j]\, e^{-\kappa|x_i - x_j|} = 0. \tag{24}$$

where $\mu_i = \frac{\kappa}{2\lambda Q_{\rm cl}(x_i)}$. Then $T(\gamma)$ may be expressed in terms of the c_i by computing the remaining integral over x (which may be done analytically).

This completes the evaluation of the distribution of χ^2. One sees that different data sets yield different $P(\chi^2)$. Therefore, it may be illustrative to consider the limit of large N, where the distribution of χ^2 assumes a more universal form.

In the limit of large N, we may put $Q_{\rm cl} \approx \bar{Q}$ and $\lambda \approx N$. We write χ^2 in a form similar to Eq. (14), but introduce a small but necessary change: $\chi^2 \approx 4N\int_{\mathbf{X}} dx\, \delta\psi^2$ where, heuristically, $\mathbf{X}$ is the region over which we may expect to find data points. We need only the size X of $\mathbf{X}$, which may be defined as $X = \frac{1}{N}\sum_i \frac{1}{Q_{\rm cl}(x_i)}$. The determinant operator is $\ell^2(-\partial_x^2 + \kappa^2)$ outside $\mathbf{X}$, and $\ell^2(-\partial_x^2 + \kappa^2(1 + \gamma))$ inside $\mathbf{X}$. Then the ratio of determinants (ignoring all but the exponential-order terms) is $D(\gamma) \approx e^{\kappa\left(\sqrt{1+\gamma}-1\right)X}$. The traces do not contribute to the exponential-order terms. Consequently,

$$\tilde{P}(\alpha) \approx e^{-\langle\chi^2\rangle\left(\sqrt{1+2\alpha}-1\right)}, \tag{25}$$

where $\langle\chi^2\rangle \approx \kappa X/\sqrt{2}$. Note that if we identify $1/\kappa$ as the effective bin width, then $\langle\chi^2\rangle$ is approximately $1/\sqrt{2}$ per bin, i.e., ≈ 0.7 per degree of freedom. We may invert the Laplace transform in Eq. (25) to obtain

$$P(z) \approx \frac{\langle\chi^2\rangle}{\sqrt{2\pi z^3}} \exp\left[\langle\chi^2\rangle\left(1 - \frac{z}{2\langle\chi^2\rangle} - \frac{\langle\chi^2\rangle}{2z}\right)\right]. \tag{26}$$

The conventional approach to statistics emphasizes the model: given a model, one calculates the likelihood of obtaining a particular data set. This likelihood is measured by the conventional χ^2. Its distribution is over (hypothetical) repeated trials of the experiment, assuming gaussian errors. In contrast, the approach presented here emphasizes the data: given a data set, one calculates the likelihood that it is described by a particular model. This likelihood is measured by χ^2; its distribution is over all possible models.

The approach presented here has two major advantages over conventional methods. First, it provides a technique for visualizing data sets, retaining all the information in the data and requiring no arbitrary choices. Second, it provides a robust measure of goodness-of-fit. Its distribution can be calculated, and so may be used for statistical analysis. The availability of a fast algorithm [5] makes computation time negligible even for large data sets. This technique should be generalizable to higher dimensions [2].

ACKNOWLEDGMENTS

TEH is supported by a Lucent Technologies Ph.D. Fellowship. I thank S. Strong and W. Bialek for useful conversations. This work is dedicated to W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling.

REFERENCES

[1] I. J. Good and R. A. Gaskins, Biometrika **58**, 255 (1971).

[2] W. Bialek, C. G. Callan, and S. P. Strong, Phys. Rev. Lett. **77**, 4693 (1996).

[3] Without such a bias, e.g., if we choose $P[Q] = 1$, the most likely Q is the solipsistic $\frac{1}{N}\sum_i \delta(x - x_i)$ [2].

[4] L. Devroye, *A Course in Density Estimation* (Birkhäuser, Boston, 1987).

[5] Eqs. (9) are solved by Newton's method, i.e., by linearizing around the solution. An $N \times N$ block of the resulting matrix equation may be put in the form $\mathbf{Au} = \mathbf{b}$, where $\mathbf{A} = \mathbf{1} + \mathbf{\Delta W}$, $\mathbf{\Delta}$ is a diagonal matrix, and $\mathbf{W}_{ij} = e^{-\kappa|x_i - x_j|}$. Note that Eq. (24) has the same form. Solving this linear equation is nominally an $\mathrm{O}(N^3)$ process. However, it is possible to do much better, because (when $x_1, \ldots, x_N$ are sorted in increasing order) $\mathbf{\Omega} = \mathbf{W}^{-1}$ is tridiagonal. Using $\mathbf{\Omega}^{-1}$ in place of $\mathbf{W}$ allows all operations to be performed in $\mathrm{O}(N)$ time, a very significant savings for large data sets. Source code may be requested from holy@puhep1.princeton.edu. Computational issues were also considered in J. Ghorai and H. Rubin, J. Stat. Comput. Simul. **10**, 65 (1979). Existence and uniqueness of a non-negative ψ_{cl} was shown in G. F. de Montricher, R. A. Tapia, and J. R. Thompson, Ann. Stat. **3**, 1329 (1975).

[6] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Urbana, 1949).

[7] P. M. Stevenson, Phys. Rev. D **23**, 2916 (1981).

[8] In Ref. [2], the smoothing parameter cannot be set until the expected value $\langle Q(x_1) \cdots Q(x_N) \rangle$ has been calculated, which requires integrating over the fluctuations and a WKB analysis. Here the fluctuations ($[2\lambda D(1)T(1)]^{-1/2}$) do not qualitatively change Figure 1; even the optimum choice for κ is changed little. Note that the choice ℓ_* in Ref. [2] is (regrettably) zero for many common distributions $\bar{Q}$.

[9] One must decide whether the λ terms are included in computing the fluctuations. The two choices yield very similar results; the version used here turns out to be somewhat simpler to implement.

[10] S. Coleman, in *Aspects of Symmetry* (Cambridge University Press, Cambridge, 1975), Chap. 7 (Appendix 1).